2009-2

Patterns of Co-Expression for Protein Complexes by Size in Saccharomyces Cerevisiae

526–532 Nucleic Acids Research, 2009, Vol. 37, No. 2 Published online 4 December 2008
doi:10.1093/nar/gkn972

Patterns of co-expression for protein complexes by size in *Saccharomyces cerevisiae*

Ching-Ti Liu[1], Shinsheng Yuan[2] and Ker-Chau Li[2,3,*]

[1]Biostatistics Department, School of Public Health, Boston University, MA 02118, USA, [2]Institute of Statistical Science, Academia Sinica, Taiwan and [3]Statistics Department, University of California, Los Angeles, CA 90095, USA

Received August 5, 2008; Revised October 30, 2008; Accepted November 18, 2008

ABSTRACT

Many successful functional studies by gene expression profiling in the literature have led to the perception that profile similarity is likely to imply functional association. But how true is the converse of the above statement? Do functionally associated genes tend to be co-regulated at the transcription level? In this paper, we focus on a set of well-validated yeast protein complexes provided by Munich Information Center for Protein Sequences (MIPS). Using four well-known large-scale microarray expression data sets, we computed the correlations between genes from the same complex. We then analyzed the relationship between the distribution of correlations and the complex size (the number of genes in a protein complex). We found that except for a few large protein complexes, such as mitochondrial ribosomal and cytoplasmic ribosomal proteins, the correlations are on the average not much higher than that from a pair of randomly selected genes. The global impact of large complexes on the expression of other genes in the genome is also studied. Our result also showed that the expression of over 85% of the genes are affected by six large complexes: the cytoplasmic ribosomal complex, mitochondrial ribosomal complex, proteasome complex, F0/F1 ATP synthase (complex V) (size 18), rRNA splicing (size 24) and H+- transporting ATPase, vacular (size 15).

INTRODUCTION

Microarray technology enables scientists to measure the expression levels of numerous genes simultaneously. Such high-throughput data sets have been a valuable resource in genome research. Many studies have been conducted to extract information from expression data, such as placing genes with similar expression profiles into the same cluster (1,2), identifying transitive genes from shortest path analysis (3) and constructing the functional module (4). A central assumption involved in these studies is the relationship between expression profile similarity and functional association. Despite many successful applications reported in the literature, to what extent, the validity of this assumption has not been fully investigated however. To approach this problem, we focused on protein complexes in yeast, a set of well-validated functional groups obtained from the Munich Information Center for Protein Sequences (MIPS) catalog (5).

We gathered four large-scale microarray data sets, Stanford cell-cycle data (2), environmental stress data (6), genetic recombination data (7) and gene deletion-perturbation data (8), and examined the correlation between genes from the same protein complex. The first question is if these within-complex correlations are higher than what would be expected for the correlations between genes that are not functionally associated. Second, the sizes of protein complexes vary greatly. Some complexes are formed by as small as two genes while other complexes can have more than 15 genes. We asked a simple question: is the strength of correlation in gene expression within a complex related with the size of the complex?

As it turns out, except for a few large protein complexes, such as the mitochondrial ribosomal proteins and the cytoplasmic ribosomal proteins, the degree of co-expression of genes from the same complex is on the average not much higher than that from a pair of randomly selected genes. We found this to be true for four well-known large expression data sets that we examined closely.

For the large protein complexes that have higher within-complex co-expression, we investigated their influence on other genes in the entire genome. We obtained strong statistical evidence suggesting that a vast majority of genes are affected by the coherent expression of these large protein complexes.

*To whom correspondence should be addressed. Tel: +1 310 825 4897; Fax: +1 310 206 5658; Email: kcli@stat.ucla.edu

MATERIALS AND METHODS

Expression data source

We analyze four publicly cDNA microarray expression data sets extensively in our study. The first data set is from Stanford's four cell-cycle experiments (2). Excluding the genes with too many missing values, 5878 genes and 73 conditions are included in total. The second data set is yeast segregation data generated by Brem *et al.* (7). In this data set, 6229 genes and 40 segregates with dye swap are included. The third data set we used is from Rosetta Compendium which contains 6280 genes and 300 expression profiles of yeast mutants (8). The last genomic expression data set was originally collected to study expression patterns responding to diverse environmental changes (6), which has 6152 genes and 173 samples.

Protein complex data source

The protein complexes were obtained from the Munich Information Center for Protein Sequences (5,9). MIPS complex catalog has tree-like structured data. To simplify the exploration of the relationship between different complexes, we excluded all 'parent' complexes with descendant complexes and used only the descendant complexes. In other words, we use the smallest indivisible units of protein complexes obtained from the MIPS complex catalog. A total of 201 protein complexes were included in our study.

Linear model

To study the influence of large protein complexes on the expression of other genes, a simple linear regression model is applied. Because the gene expression profiles within these protein complexes are tightly clustered, we used the mean expression profile as the consensus profile for each complex. The mean expression profile here means the average for each specific condition point in the expression profile. More precisely, suppose there are K genes in a complex whose expression are profiled for a total of 'd' conditions and are denoted by d-dimensional vectors $g_1, \ldots, g_K$. Then the consensus profile is a d-dimensional vector $\mathbf{C} = (g_1 + \cdots + g_K)/K$. We regressed each gene profile in the genome against the consensus profile. Specifically, for each gene i and complex j,

$$G_i = \beta_0 + \beta_1 C_j + \varepsilon_i, \quad \mathbf{1}$$

where G_i indicates the ith gene's expression profile and C_j indicates the jth complex's expression consensus profile. The gene's expression level is said to be associated with a selected complex if this fitness of model is better than preselected criterion. Specifically, the criterion is determined as the follows:

(i) Randomly permute two gene expression levels, say g_1 and g_2;
(ii) Use the permuted gene expression level to fit a simple linear regression model $g_1 = \beta_0 + \beta_1 g_2 + \varepsilon$;
(iii) Assess and record the goodness of fit for the model. Here, we use coefficient of determination or R^2;
(iv) Repeat all the procedures and collect all the R^2 for each iteration; and
(v) Sort R^2 and calculate its 95th percentile, which is used as the threshold.

In other words, this criterion is chosen to guarantee its superiority to at least 95th percentile of permuted data. We applied this procedure to four different expression data sets for deciding the threshold for each data set.

Evaluation of significance

The results of the estimation of the linear model can help us assess if a protein complex may have any impact on the expression of a specific gene. A gene's expression level is said to be associated with a selected complex if the goodness of fit of the model is better than the prespecified criterion defined in the previous section. Following this procedure, we obtained the total number of genes in the genome whose expression levels can be explained by the consensus profile of a protein complex. Then we ask if this number, denoted by 'T', is significantly higher than what would be expected by pure chance, should there be no links between the consensus profile and the genome-wide expression profiles. To create the no-link situation by simulation, we conduct random permutation in the following way.

Suppose the microarray profiling is done under a total of d conditions labeled by 1, 2, ... d. The output gene profiles for the entire genome of N genes can be represented by a matrix M of d columns and N rows, each row representing one gene expression profiles. The consensus profile of a given complex under study is a d-dimensional vector $\mathbf{C} = (c_1, \ldots c_d)$'. We now randomly permute the columns of M to break the link between the full genome profiles and the consensus profile. We then regress each row of the permuted matrix M on the consensus profile **C** to see if the fit is significant (R-squared better than pre-determined threshold). Because the link between **C** and the M is broken by randomization now, the significant fit is deemed as being obtained by pure chance. We recorded the number of significant rows, denoted by T_r and compare it with the number T that we have at hand from the true data. We repeat the randomization $n = 1000$ times and record the proportion of times when $T_r \geq T$. This serves as the P-value for evaluating the significance of the observed number of genes associated to the given protein complex.

RESULTS

Expression analysis of protein pairs

In addition to the protein complexes in yeast, we also constructed a negative control set by gathering all proteins from different cellular components to form unrelated pairs as done in (10). Protein localization provides important information for elucidating Eukaryotic protein function (11). Presumably, proteins in different cellular components are less likely to interact with each other and the biological processes they participate are different in general. Thus, this negative set of unrelated pairs provides a contrast to

the within-complex pairs of functionally related genes. To control the ill-effect of outliers in the data, we had conducted normal-score transformations for each gene's expression level in the data preprocessing step as was done in (12).

We compute Pearson correlation for each pair, using four gene expression data sets. Distributions of correlations are plotted in the left panel of Figure 1. We denote the protein pairs from different cellular components as unrelated pair (abbreviated as 'unrel') while the pairs within the same protein complex as related pair (abbreviated as 'rel'). The terms 'cc', 'yg', 'rst' and 'st1' represent four different data sets: cell cycle, segregation genetics, rosetta and stress data, respectively. In the left panel of Figure 1, correlation distribution for protein complex pair is plotted using the solid line while correlation of protein pairs with different cellular localizations is plotted using the dash line.

As expected, the four curves which represent the results from unrelated pairs are seen to center around 0 symmetrically. This indicates the lack of coordinate expression overall for protein pairs from different cellular localizations. In contrast, the other four curves for within-complex pairs are shifted to the right. In other words, this figure shows that within-complex pairs overall have significantly higher correlation than the unrelated protein pair. Unfortunately, this turns out to be a rather misleading impression.

Pattern of co-expression by complex size

The number of genes in a protein complex varies substantially, ranging from 2 to 81 in our data set. Figure 2(a) shows the histogram of the sizes of protein complexes. For instance, around 83.6% (or 93.0%, respectively) of protein complexes have size <10 (or 15 respectively). However, by the sheer number of combinations alone, large complexes account for the majority of within-complex gene pairs. For instance, there are in total 2371 gene pairs from protein complexes of size 14 or less; but there are in total 7964 gene pairs from protein complexes of size 15 or more. This explains why the related-pair curves shown in the left panel of Figure 1 look more like the distribution for gene pairs from large complexes only on the right panel.

The right panel of Figure 1 depicts the comparison of correlation distribution for small complexes (size 14 or less) and that for large complexes (size 25 or more). The mode of the correlation distribution representing small complexes, in each of the four expression data sets, is now seen to shift back toward 0 substantially (from 0.169, 0.298, 0.082, 0.189 to 0.435, 0.827, 0.496, 0.898, respectively).

A possible explanation for the much reduced correlation in gene expression is that the transient type of interaction between proteins may be more common for smaller complexes. Such interactions tend to take place briefly only under very specific cellular conditions. Thus, even if there is a co-regulation at the gene expression level under specific conditions, such short-lived pattern of coordinate expression is harder to detect by statistical correlation which is formulated for depicting the global coordinate pattern across all conditions.

We conduct a more detailed study on how the correlation pattern depends on the size of a protein complex. For each complex of size n, we obtain the median of the $\binom{n}{2} = n(n-1)/2$ correlations computed for each gene pair within the complex. Then the median correlations for complexes of the same size are put together and a boxplot is drawn to show the distribution by size; see Figure 2(b).

To measure the significance of co-expression, we randomly permute the expression levels of gene pairs and compute its correlation. Then the significant level of

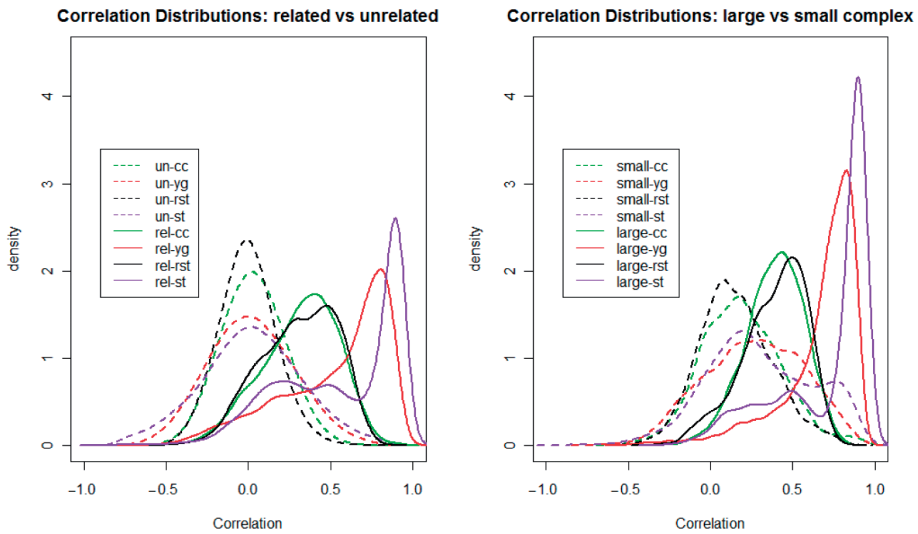

Figure 1. Comparison of correlation distributions for protein pairs with respect to functional association (shown in left panel) and complex size (shown in right panel). The terms 'cc', 'yg', 'rst' and 'st1' represent four different data sets: cell cycle, segregation genetics, rosetta and stress data, respectively. Protein complex pairs are abbreviated as 'rel' and unrelated pairs are abbreviated as 'unrel'.

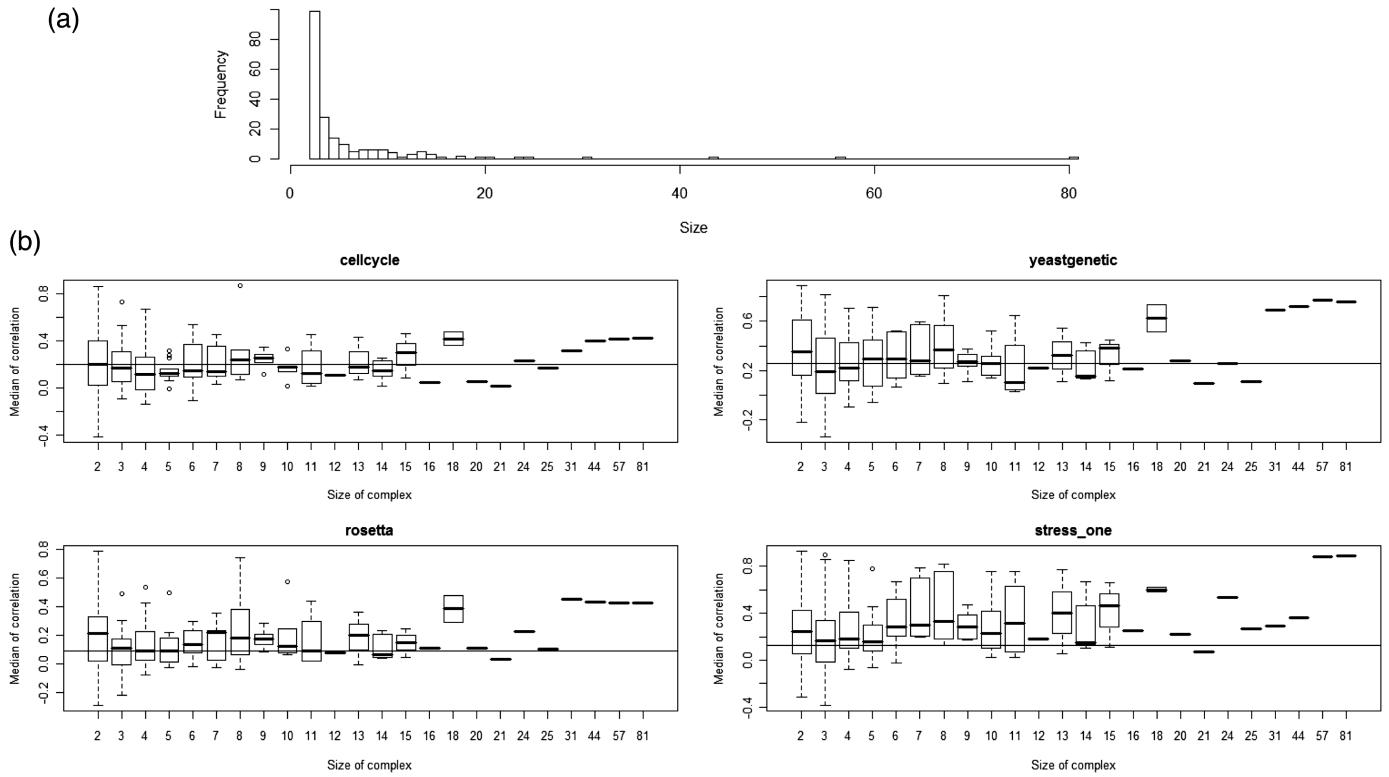

Figure 2. (**a**) Histogram of protein complex size. (**b**) Boxplots are used to sort out the relationship between the size of protein complex and the median correlation for gene pairs within a complex.

correlation is determined from permuted correlation distribution. This procedure is carried out for each of the four expression data sets. The horizontal line is placed in Figure 2(b) to indicate 95% significant level obtained from permutation. To determine if the co-expression level for protein complexes is higher compared to the co-expression level of any random gene pairs 95% significant level is used as a threshold. Thus, for each step, we simply generate the expression profiles for gene pair and calculate its correlation. Repeat this procedure 10 000 times and then the 95th percentile from the simulated correlation distribution is found to be 95% significant threshold in this analysis.

Ribosome

The largest four complexes are mitochondrial ribosomal small subunit (size 31), mitochondrial ribosomal large subunit (size 44), cytoplasmic ribosomal small subunit (size 57) and cytoplasmic ribosomal large subunit (size 81), all showing very strong within-complex co-regulation. We further examined the correlation for the gene pair between complexes. We found that the two cytoplasmic ribosome complexes have very high cross-complex correlation and so do the two mitochondria complexes (Figure S1, Supplementary Data).

rRNA splicing (size 24)

This complex also has high within-complex correlation in gene expression. Because rRNA and the ribosome proteins are both critical components of the translation machinery, we may speculate a likely co-regulation between rRNA splicing complex and any of the ribosomal complexes in expression. However, the results between rRNA splicing complex and mitochondrial ribosomal complex are negative (Figure S2, Supplementary Data).

Complexes of size 18 and size 15

There are two complexes size 18: 19/22S regulator complex and F0/F1 ATP synthase complex. Both have very high within-complex co-expression patterns.

There are three complexes with size 15: SAGA complex, 20S proteasome, and H+-transporting ATPase vacular complex. A closer inspection shows that the correlation for gene pairs from SAGA complex is not as strong as those from the other two complexes.

We noticed that the highly co-expressed genes in 20S proteasome complex of size 15 have high cross-complex correlation with genes from the 19/22S regulator complex of size 18. According to protein complex catalog in MIPS, these two protein complexes form the 26S proteasome complex (of size 36) along with three additional proteins *RPN4*, *DOA4* and *YTA7*. However, the expression profiles of *RPN4, DOA4* and *YTA7* are not correlated to these two complexes at all (Figure S3M Supplementary Data).

From our discussion, we may conclude that complexes with larger size are more likely to have co-expressed genes. Our findings are in line with the results

of transient/permanent complex partition (10). This suggests that the complex size is an important confounding factor of strength of co-expression and we have to take more caution when using the approach of 'guilt by association'.

Highly co-regulated large complexes

By combining the co-regulated complexs discussed above, we end up with six large complexes: mitochondrial ribosomal unit (size 31 + 44), cytoplasmic ribosomal unit (size 57 + 81), proteasome complex (size 18 + 15), F0/F1 ATP synthase (complex V) (size 18), rRNA splicing (size 24) and H + - transporting ATPase, vacuolar complex (size 15) (Table 1).

Co-regulation between complexes

The mean profile is used to represent the consensus profile for each complex. From the correlation matrix between the consensus profiles of complexes in Table A1 shown in Supplementary Data, we found strong correlation between mitochondrial ribosomal complex, ATP synthase complex and proteasome complex. Strong co-expression is also observed among cytoplasmic ribosomal complex, rRNA splicing complex and H + -transporting complex. The correlation between rRNA splicing and mitochondrial-group protein complexes and that between ATP synthase complex and cytoplasmic-group protein complexes are mild. However, there is no correlation between cytoplasmic ribosomal complex and mitochondrial ribosomal complex. We may put these six complexes into two classes: one consisted of cytoplasmic ribosomal complex, rRNA splicing complex and H + -transporting complex; the other consisted of mitochondrial ribosomal complex, ATP synthase complex and proteasome complex.

Table 1. List of highly co-expressed complexes with size ≥15

Complex name	Complex size
Cytoplasmic ribosomal complex	138
Mitochondrial ribosomal complex	75
Proteasome complex	33
rRNA splicing complex	24
F0/F1 ATP synthase (complex V)	18
H + -transporting ATPase, vacular	15

Impact on other genes

Proteins in the same complex carry out biological processes together. Treating a group of associated proteins as a functional modular, we are interested in how well the regulatory activities of genes can be explained by the identified large and highly co-expressed complexes. We applied a simple linear regression to study the influence of protein complexes on the expression of other genes. A gene's expression level is said to be associated with a complex if its goodness of fit for the model is better than the preselected criterion. The results recording the number of genes significantly impacted by the identified protein complex are shown in Table 2

Table 2. The number of genes whose expression level can be explained by complex

	Mitochondrial	Cytoplasmic	rRNA splicing	Proteasome	ATP synthase	H + -transporting
Cellcycle (5878)						
Mitochondrial	2373 (0.026)	677 (0)	1040 (0)	1901 (0)	1456 (0)	1349 (0)
Cytoplasmic		1859 (0)	1141 (0)	819 (0)	805 (0)	935 (0)
rRNA splicing			2545 (0)	1200 (0)	1049 (0)	1022 (0)
Proteasome				2802 (0)	1568 (0)	1490 (0)
ATP synthase					2343 (0)	1680 (0)
H + transporting						2490 (0)
Genetic (6229)						
Mitochondrial	3349 (0)	2116 (0)	1985 (0)	2353 (0)	1652 (0)	1738 (0)
Cytoplasmic		4008 (0)	3241 (0)	2628 (0)	1635 (0)	2089 (0)
rRNA splicing			3933 (0)	2505 (0)	1551 (0)	2019 (0)
Proteasome				3734 (0)	1471 (0)	1840 (0)
ATP synthase					2601 (0)	1304 (0)
H + transporting						3169 (0)
Rosetta (6283)						
Mitochondrial	4001 (0)	2791 (0)	2723 (0)	2094 (0)	2302 (0)	2187 (0)
Cytoplasmic		4379 (0)	3101 (0)	1923 (0)	2468 (0)	2503 (0)
rRNA splicing			4237 (0)	1868 (0)	2433 (0)	2391 (0)
Proteasome				2819 (0)	1539 (0)	1510 (0)
ATP synthase					3432 (0)	1933 (0)
H + transporting						3508 (0)
Stress (6152)						
Mitochondrial	3083 (0.024)	3258 (0)	3283 (0)	2274 (0)	3063 (0)	3123 (0)
Cytoplasmic		5184 (0)	4795 (0)	2924 (0)	4228 (0)	4592 (0)
rRNA splicing			5129 (0)	2864 (0)	4224 (0)	4438 (0)
Proteasome				3554 (0)	2579 (0)	2836 (0)
ATP synthase					4642 (0)	4083 (0)
H + transporting						5053 (0)

Table 2 shows the number of genes associated with each complex (given in the diagonal cells) and with two complexes (given in the off-diagonal cells). The number in the parenthesis is the *P*-value obtained from simulation. The detail information is shown in 'Materials and methods' section. Take the cell-cycle data for example. There are 2373 genes whose expression profiles are associated with mitochondrial ribosome complex and the *P*-value is.026. The number of genes associated with both the proteasome complex and mitochondrial complex is 1901.

The cytoplasmic ribosomal complex appears to have the largest number of expression-associated genes across four different data sets. It is followed by rRNA splicing complex. The numbers of genes associated with the rRNA splicing and the cytoplasmic complex, especially in the yeast genetic and stress data, are also very large. This observation is consistent with the high correlation between the consensus profiles of these two complexes.

To summarize, the expression profiles of 5085 (86.5%), 5992 (96.2%), 6213 (98.9%) and 6072 (98.9%) genes are related to at least one of the six large and coherently expressed protein complexes in the cell cycle, segregation, rosetta and stress data set, respectively. This well illustrates how extensive the impact of these six selected complexes is in many biological processes. Specifically, using cell-cycle data set, we identified 2373 genes related to mitochondrial ribosomal complex. Among these genes, we conducted gene set enrichment analysis on Gene Ontology and found that many enriched gene ontology terminology are related to mitochondria. For example, we observed the enriched term such as protein folding (*P*-value = 0.00031), mitochondrion organization (8.74E-12), aerobic respiration (1.53E-8), mitochondrial transport (1.33E-6), mitochondrial transport (1.33E-6), protein targeting to mitochondrion (2.50E-5).

DISCUSSION

Biological processes are regulated or carried out by groups of molecules usually referred to as functional modules. Analyzing different functional module separately may obscure the complexity of cellular machinery. To shed some light on this issue, we have explored the multi-protein complexes by integrating data from mRNA microarray expression, protein localization and protein complex information.

To study the relationship between expression profile similarity and functional association, some authors have reported co-expression modules that had enrichment in rRNA processing, protein synthesis and the ubiquitin pathway gene ontology terms (13). However, in the data we investigated, not all proteins from the same protein complex exhibit common co-expression patterns. Our results indicate that the correlation in the expression profiles of proteins from the same complex is stronger for larger-sized complexes while for complexes of small sizes, it varies a lot.

Some large complexes are found to be lack of the co-expression patterns shown in the six large and coherently-expressed complexes we identified. These complexes may have more complicated expression structure. For example, they may be decomposed into subgroups of genes within which the expression patterns are similar. Specifically, we consider the replication protein complex as an example. The replication protein complex has 20 proteins in the complex. The average of the within complex correlations is only 0.115 in cell-cycle data. The values in other datasets are also low. We applied the centroid hierarchical clustering method to these 20 expression profiles. A clear co-expression pattern was observed among subgroup of replication protein complex. The clustering result is shown in Figure 3.

From Figure 3, we can see that replication protein complex can be divided into three subclusters. One group (abbreviated as MCM-group) is *MCM2, MCM3, MCM4 (CDC54)*, *MCM5 (CDC46)*, *MCM6*, *MCM7 (CDC47)*, another group (abbreviated as *ORC*-group) is *ORC1*, *ORC2*, *ORC3*, *ORC4*, *ORC5*, *ORC6* and the remaining genes, *DPB11*, *DPB3*, *HYS2*, *CDC2*, *POL32*, *CDC45*, *DPB2* and *POL2*, form the third group. The average of the within-group correlations between the expression profiles of genes from the same group is 0.633, 0.173 and 0.448 for *MCM*-group, others-group and *ORC*-group, respectively. However, the correlations for gene pairs from different groups is low. For example, the average is –0.054 in the correlation between genes in *MCM*-group and genes in *ORC*-group. This example well illustrates the existence of subcluster in replication protein complex can lower the overall correlations between genes in a complex.

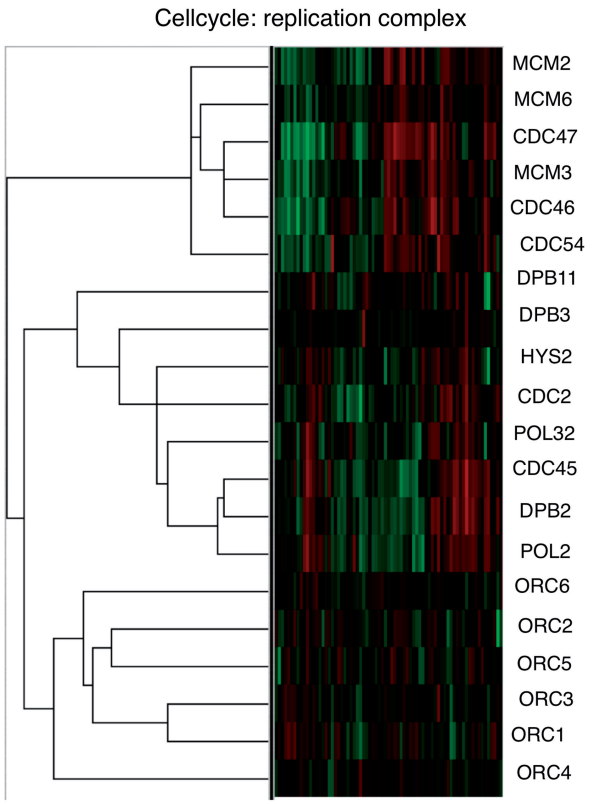

Figure 3. Hierarchical clustering with centroid method. The cluster result using cell-cycle expression data clearly shows three subclusters in the replication protein complex.

This finding is actually supported by other research results such as references 14 and 15 or documents on Saccharomyces Genome Database (SGD and its URL is http://www.yeastgenome.org). The origin recognition complex (ORC-group) is essential for MCM binding to chromatin (14) and permits the loading of other replication factors onto origin DNA (15). The minichromosome maintenance proteins (MCM-group or *MCM2-7* family) is involved in the initiation of DNA replication.

Another large complex in Figure 2(b) has poor correlation. This complex of size 21 represents Kornberg's mediator (SRB) complex. The insignificance of the overall correlations may be due to the more complex gene regulatory mechanism. For example, although this complex was assigned as a terminal node according to the curated information by MIPS, In 1998, Lee and Kim (16) pointed out that the mediator can be dissociated into two stable subcomplexes, the Rgr1-containing subcomplex and the Med6-containing subcomplex.

In this study, we have investigated the protein complex by analyzing the correlation structure of their protein member. Alternatively, one of the reviewers suggested considering the 'density' of co-expression network used in the co-expression network (17). By the 'density' of co-expression, these network researchers consider absolute value of correlation rather than correlation itself. Following this alternative measure of co-expression for each complex, we carried out the same analysis as in Figure 2(b) and reported the results in Figure S4. As expected, Figure 2(b) and Figure S4 appeared similar to each other and we did not find noticeable differences between the absolute values of correlations and the direct correlation. One reason may be because in our protein complex study, genes tend to be positively co-regulated within one complex (especially if complexes are structural, rather than regulatory) in order to function coherently. This is an important aspect about protein complexes, which is somewhat different from pathway studies where genes could be regulated in a compensated manner to serve as feedback control, thus generating possible negative correlations.

Characterization of protein interaction is an important issue from many aspects, such as drug design. In our study, we have studied known protein complexes by integrating independent but related data. This integrative study has uncovered the size effect of protein complex which may work differ spatially or temporally. It provides a new perspective on the study of protein complex or more general protein–protein interactions. Additionally, our study may also provide insights to study co-regulation between complexes and the construction of the eigengene networks (18).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

NSF grants DMS0201005, DMS0406091 and DMS-0707160 (to K.-C.L.); MIB, Institute of Statistical Science, Academia Sinica (grants NSC95-3114-P-002-005-Y and NSC97-2627-P-001-003 to K.-C.L. and S.Y.). Funding for open access charge: MIB, Institute of Statistical Science, Academia Sinica.

Conflict of interest statement. None declared.

REFERENCES

1. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
2. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M., Brown,P., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3279.
3. Zhou,X., Kao,M.C. and Wong,W.H. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl Acad. Sci. USA*, **99**, 12783–12788.
4. Zhang,B. and Horvath,S (2005) A general framework for weighted gene co-expression network analysis. *Statistical Appl. Genet. Mol. Biol.*, **4**, 17.
5. Mewes,H.W., Frishman,D., Gueldner,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Munsterkotter,M., Rudd,S. and Weil,B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.
6. Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.
7. Brem,R.B., Yvert,G., Clinton,R. and Kruglyak,L. (2002) Genetic dissection of transcriptional regulatioin in budding yeast. *Science*, **296**, 752–755.
8. Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
9. Mewes,H.W., Frishman,D., Mayer,K.F.X., Munsterkotter,M., Noubibou,O., Pagel,P., Rattei,T., Oesterheld,M., Ruepp,A. and Stumpflen,V. (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.*, **34**, 169–172.
10. Jansen,R., Greenbaum,D. and Gerstein,M. (2002) Relating whole-genome expression data with protein–protein interactions. *Genome Res.*, **12**, 37–46.
11. Kumar,A., Agarwal,S., Heyman,J.A., Matson,S., Heidtman,M., Piccirillo,S., Umansky,L., Drawid,A., Jansen,R., Liu,Y. *et al.* (2002) Subcellular localization of the yeast proteome. *Genet. Dev.t*, **16**, 707–709.
12. Li,K.-C. (2002) Genome-wide coexpression dynamics: theory and application. *Proc. Natl Acad. Sci. USA*, **99**, 16875–16880.
13. Carlson,M.R.J., Zhang,B., Fang,Z., Mischel,P.S., Horvath,S. and Nelson,S.F. (2006) Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics*, **7**, 40.
14. Madine,M.A., Swietlik,M., Pelizon,C., Romanowski,P., Mills,A.D. and Laskey,A. (2000) The roles of the MCM ORC, and Cdc6 proteins in determining the replication competence of chromatin in quiescent cells. *J. Struct. Biol.*, **129**, 198–210.
15. Kearsey,S.E. and Labib,K. (1998) Mcm proteins: evolution, properties, and role in DNA replication. *Biochem. Biophys. Acta*, **1398**, 113–136.
16. Lee,Y.C. and Kim,Y.-J. (1998) Requirement for a functional interaction between mediator components Med6 and Srb4 in RNA polymerase II transcrippition. *Mol. Cell. Biol.*, **18**, 5364–5370.
17. Horvath,S. and Dong,J. (2008) Geometric interpretation of gene coexpression network analysis. *PLoS Computat. Biol.*, **4**, e1000117.
18. Langfelder,P. and Horvath,S. (2007) Eigengene networks for studying the relationships between co-expression modules. *BMC Syst. Biol.*, **1**, 54.